VERZEO MAIN PROJECT

By:- ALOK KUMAR
Student in PRO-DS
alokanshu16789@gmail.com

Mentor:- Smit Sha

Objectives

The main task is to find cluster of the at least 5 countries which are in dire need of aid.

APPROACH:

I will analyze first K-Means after then Hierarchical clustering by passing all the important steps.After viewing the result of both the clustering I will choose the best between both to get the required countries.

Description of the dataset.

	Column Name	Description
0	country	Name of the country
1	child_mort	Death of children under 5 years of age per 1000 live births
2	exports	Exports of goods and services per capita. Given as %age of the GDP per capita
3	health	Total health spending per capita. Given as %age of GDP per capita
4	imports	Imports of goods and services per capita. Given as %age of the GDP per capita
5	Income	Net income per person
6	Inflation	The measurement of the annual growth rate of the Total GDP
7	life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
8	total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
9	gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

STEPS and its RESULT

Steps: Finding info() on dataset.

Result: No null value.

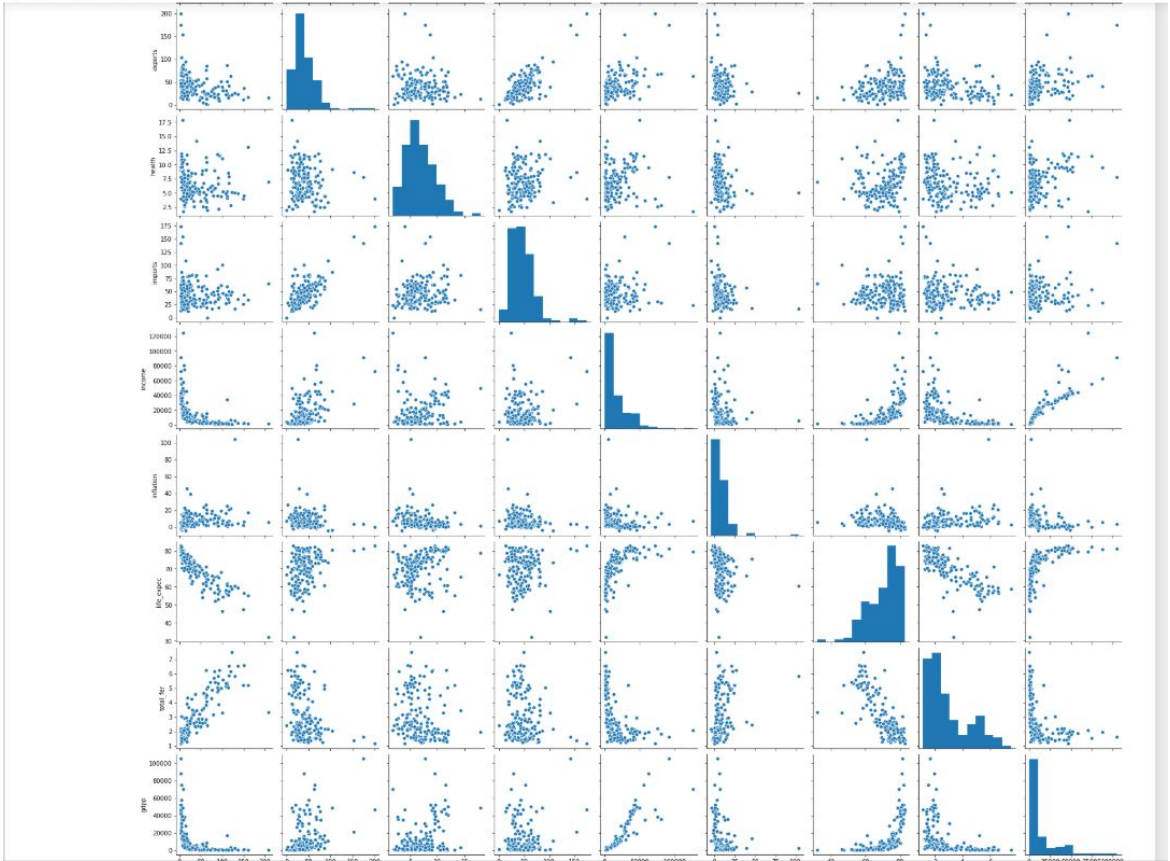
Conclusion:There is no null value so there is no need to exempt any data row and now we can proceed for K means clustering.

Steps: Plotting the Boxplot.

Result: Refining the data by choosing outliner.

Univariate and bivariate analysis

Pairplot



Observations

1.Higher the income,higher the gdp,higher the life expectancy

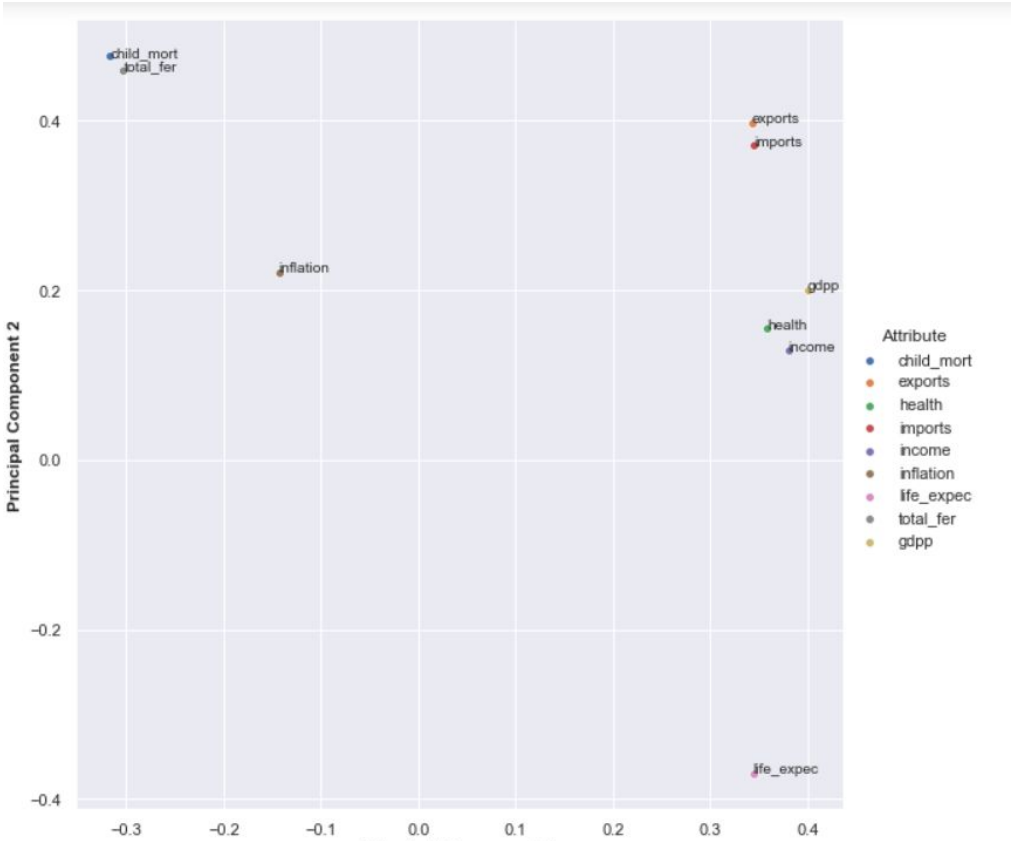
Means people in higher income countries tend to live longer.

2.Higher the income lower the child mortality.

3.Lower the inflation lead to good health.

Note: converted exports, imports and health spending percentages to absolute values.

Plotting the dataframe by the PCA components.



Inferences

1.Life expectancy, income, gdpp and health are very well explained by PC1.

2. Imports and exports are well explained by both the components PC1 and PC2.

3. child mortality and total fertility are well explained by PC2.

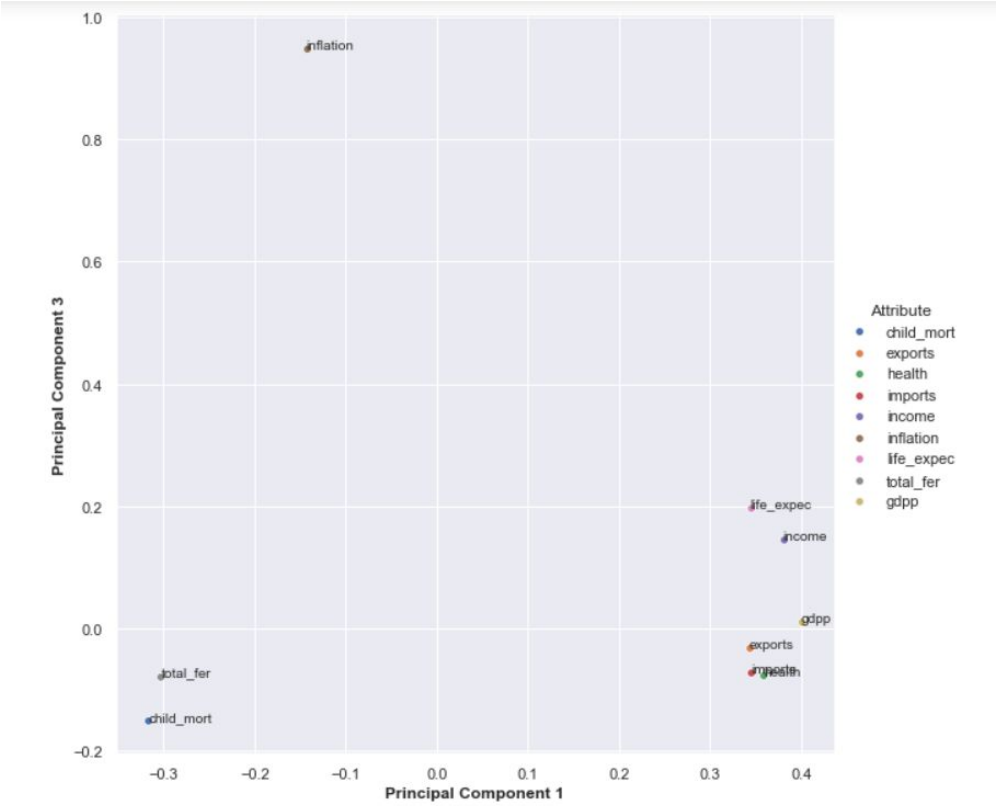
Inflation is well explained by PC3

Conclusion:

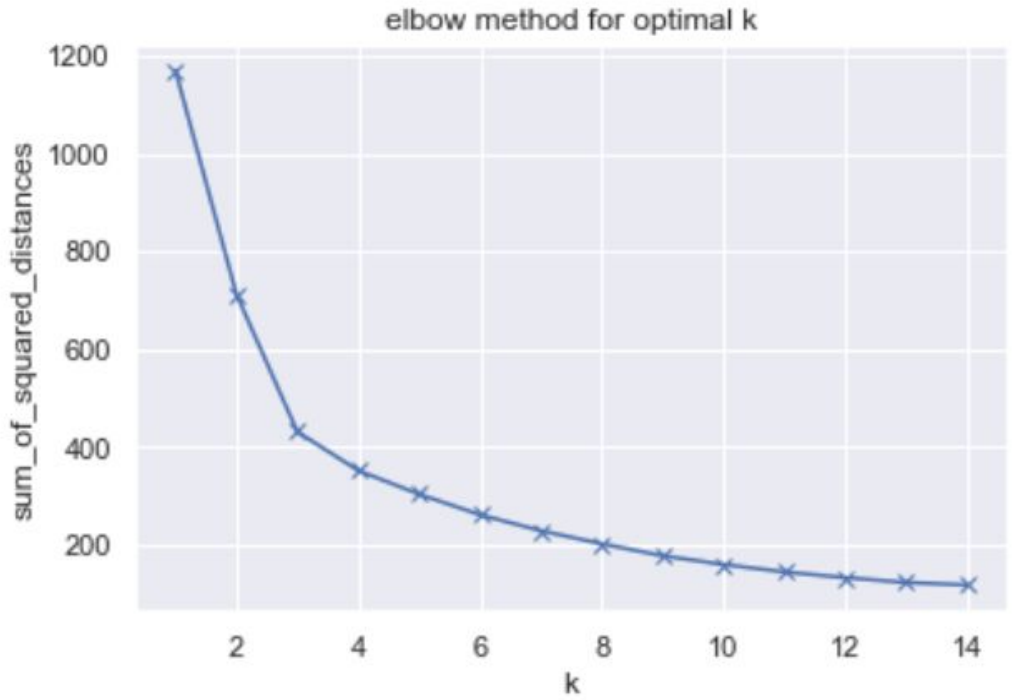
So I choose PC1 ,PC2

And PC3 as PCA

components.

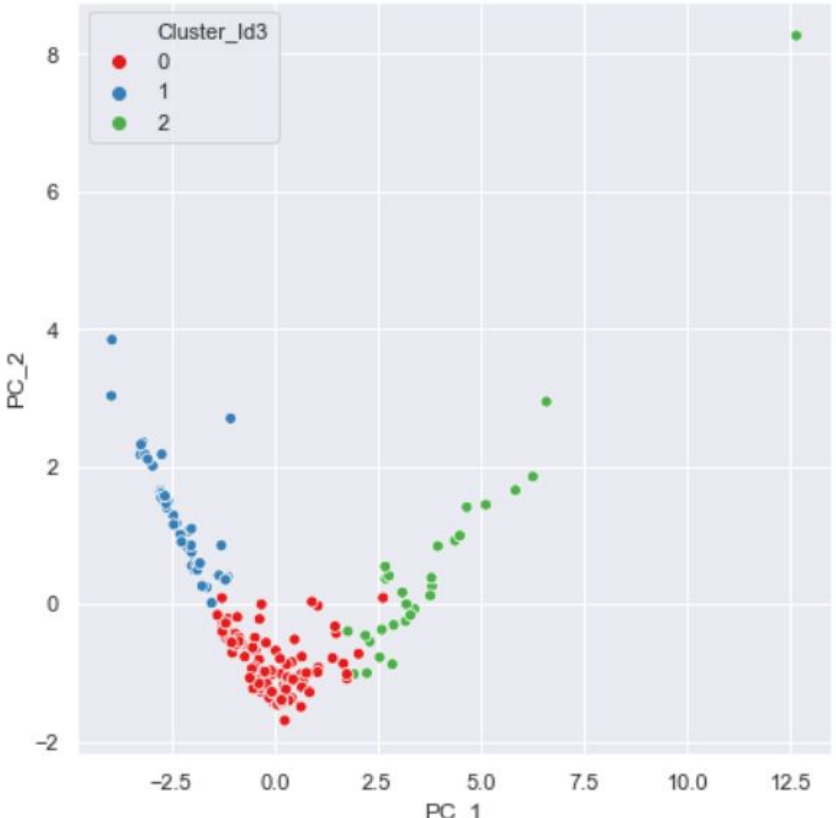


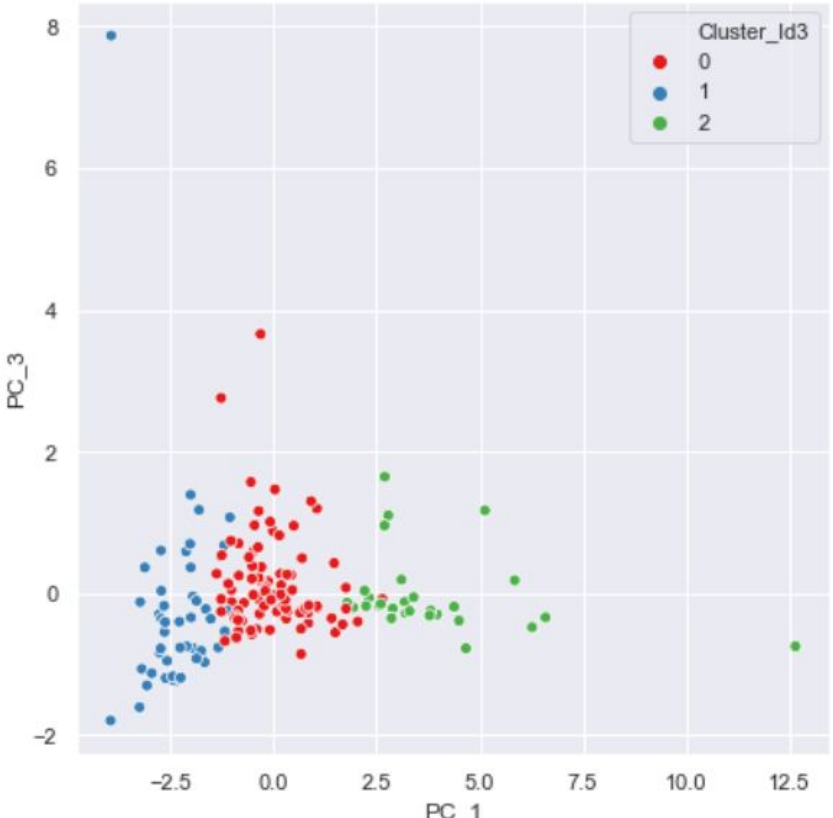
An elbow plot to determine value of K



K=3(elbow)

Scatter plot on Principal components to visualize the spread of the data





Conclusion

Intradistance between the clusters is not more i.e they are not randomly far from each other so k=3 is good choice.

K-Means clustering on k=3

	Cluster_Id3	Child_Mortality	Exports	Imports	Health_Spending	Income	Inflation	Life_Expectancy	Total_Fertility	GDPp
0	0	20.5644	3817.8411	3909.0991	516.9122	13789.8889	7.1645	73.3711	2.2249	7664.1333
1	1	91.6104	879.0635	827.0288	114.8218	3897.3542	11.9111	59.2396	4.9921	1909.2083
2	2	4.9310	29429.0552	24439.2586	4291.0655	49482.7586	2.8629	80.5483	1.8086	47710.3448

Observations

Cluster 1 has lowest GDPP.

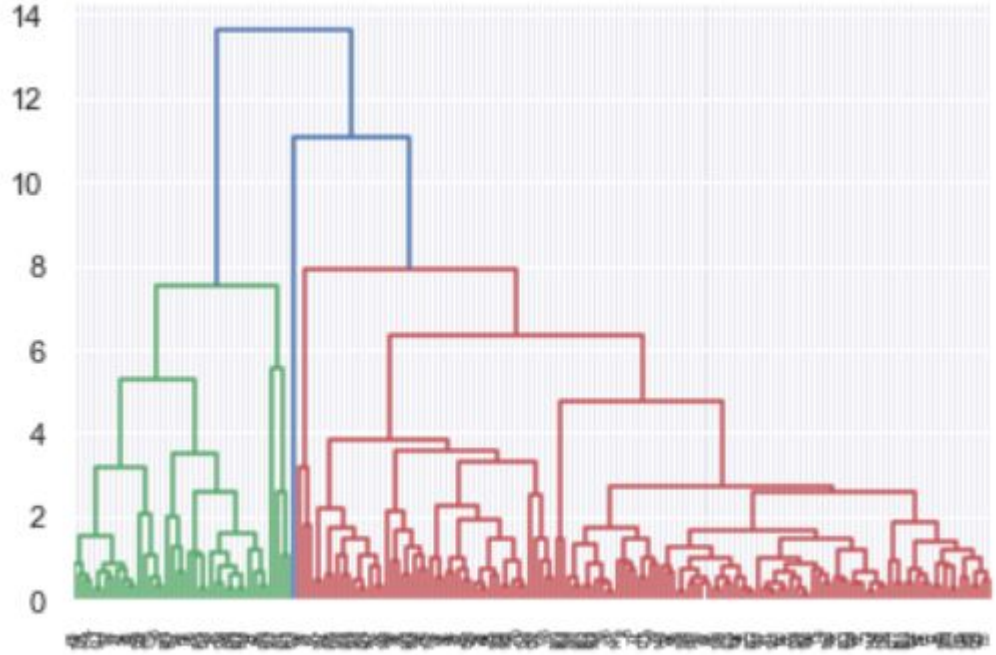
Cluster 1 has highest child_Mortality, Inflation and Total_fertility

Cluster 1 has lowest imports,exports and life_expectancy and income

so the countries within cluster 1 need fund more than countries of cluster 0 and 2.

** So as of now 48 countries of cluster 1 is chosen**

HIERARCHICAL CLUSTERING¶

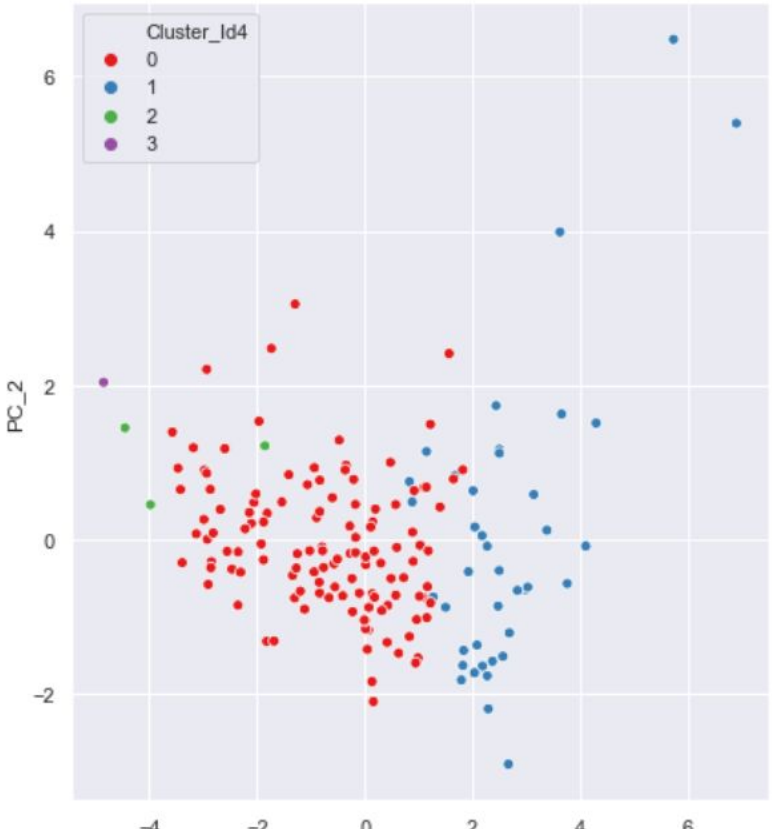


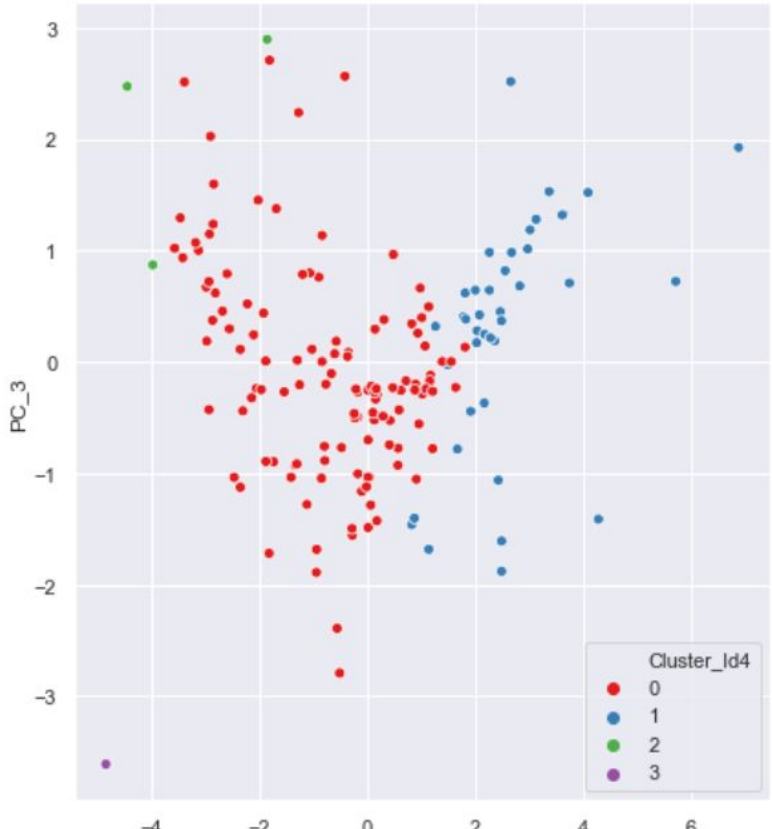
HECK FOR CUTTING TREE

Find the longest vertical line which should not be cut down by horizontal line and then count the number of intersection of that cut vertical and imaginary horizontal line.

So I cut the tree at the height of approx 7 and got 4 cluster.

Scatter plot on Principal components to visualize the spread of the data

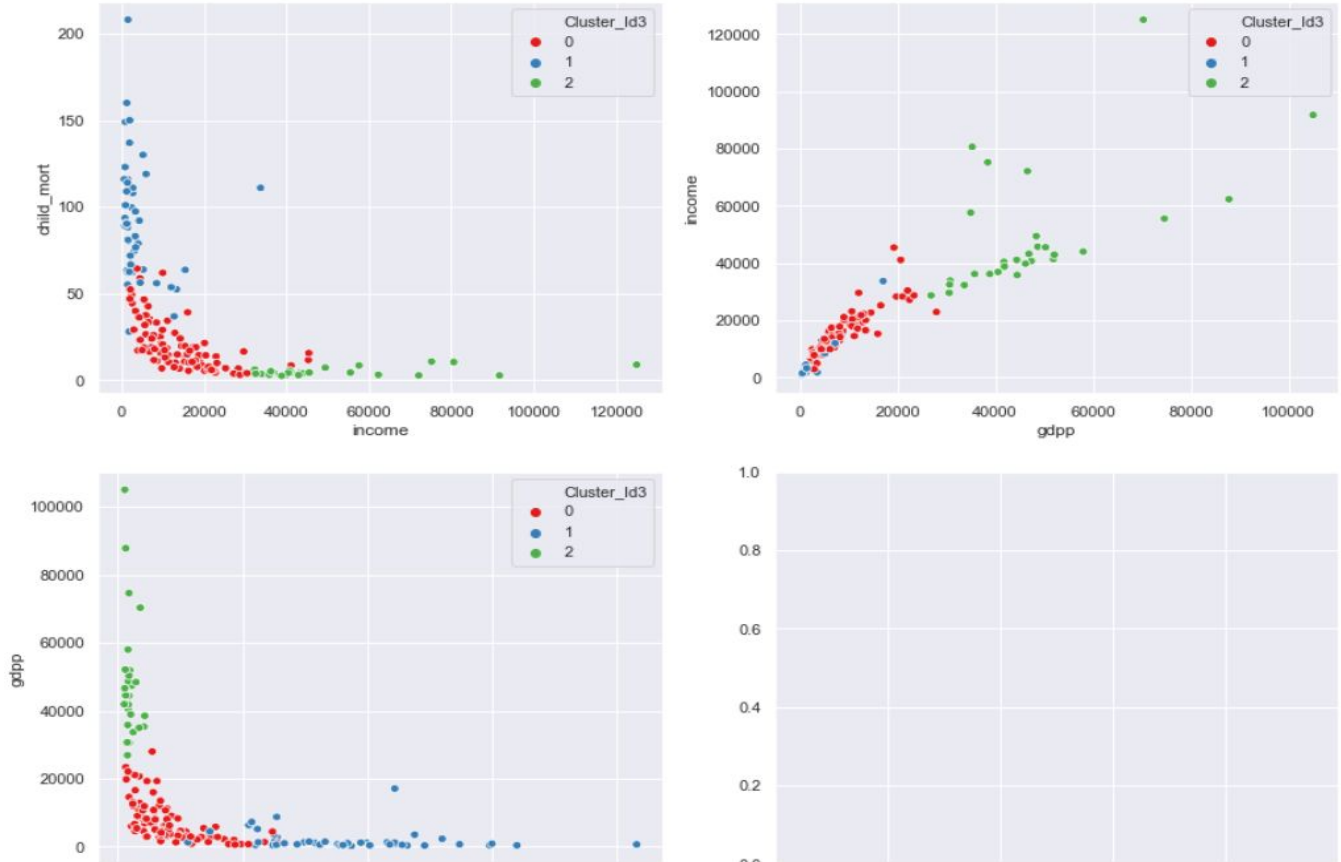




Conclusion

since cluster is well organised i.e they are not far from each other so we will choose no. of cluster=4.

Scatter plot on Original attributes to visualize the spread of the data



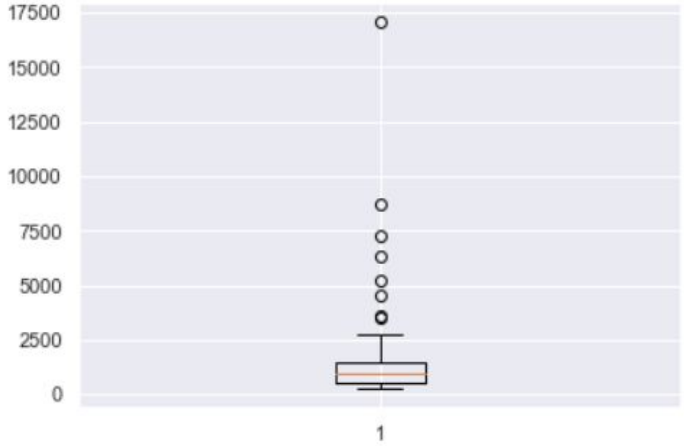
Conclusion

From the above graph and the list of countries with respective clustering i came to know that in hierarchical clustering model there is ambiguity ex.. serial no 161 country name=Uzbekistan with 1380 gdpp should lie within cluster 1 instead of cluster 0 because it has almost same gdpp of countries 165 and 166 named Yemen and Zambia with cluster 1 same is the case with 164 Vietnam and many more.

So I finalise K- Means clustering for choosing the country.

162	Vanuatu	29.2000	1384.0200	1565.1900	155.9250	2950	2.6200	63.0000	3.5000	2970	0
163	Venezuela	17.1000	3847.5000	2376.0000	662.8500	16500	45.9000	75.4000	2.4700	13500	0
164	Vietnam	23.3000	943.2000	1050.6200	89.6040	4490	12.1000	73.1000	1.9500	1310	0
165	Yemen	56.3000	393.0000	450.6400	67.8580	4480	23.6000	67.5000	4.6700	1310	1
166	Zambia	83.1000	540.2000	451.1400	85.9940	3280	14.0000	52.0000	5.4000	1460	1

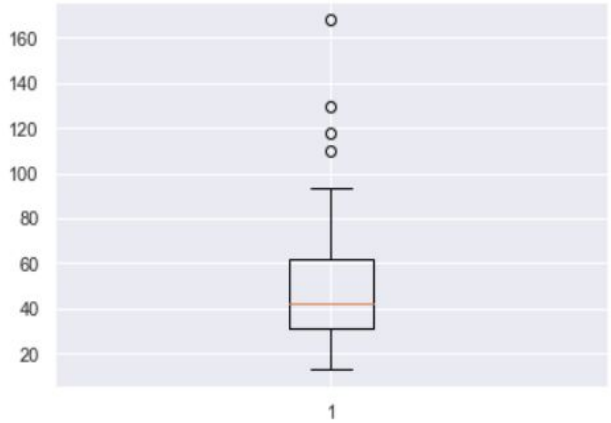
Outlier Treatment



In [324]:
```
# we are observing that some of the good(higher gdpp) countries comes to the list so we are going to remove these countries
df_final2 = df_final1[df_final1['gdpp']<2500]
df_final2.shape
```

Out[324]: (39, 11)

For health



In [328]:
```
# we are observing that some of the good health expanding countries comes to the list so we are going to remove these countries.
df_final3 = df_final2[df_final2['health']<100]
df_final3.shape
```

Out[328]: (35, 11)

For inflation

No outlier no need to remove any data regarding inflation

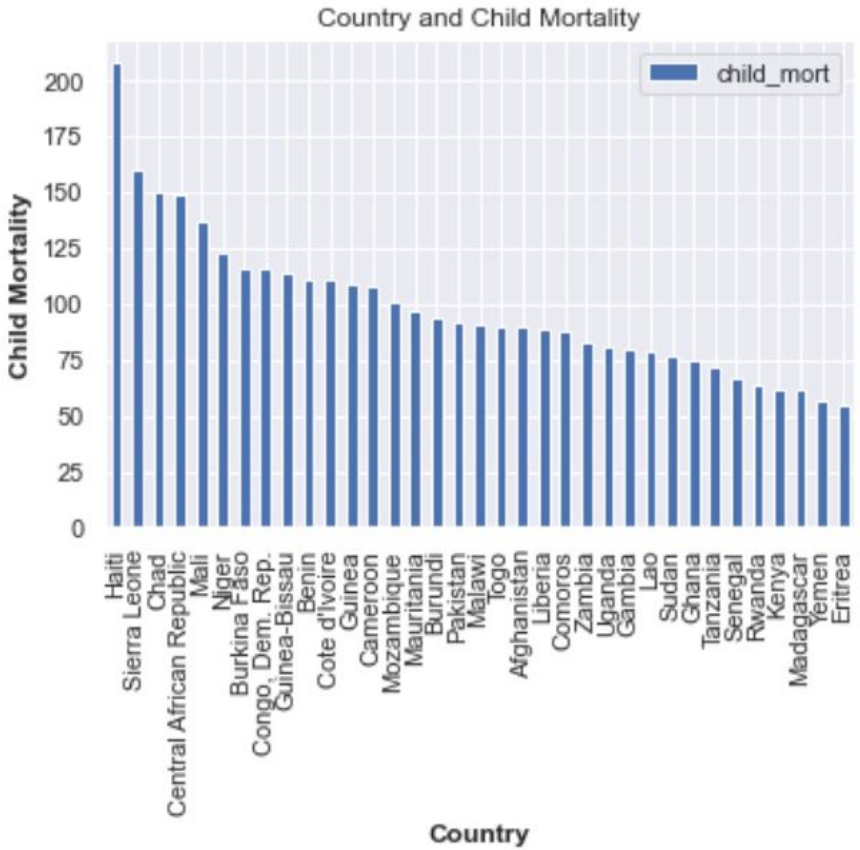
FINAL COUNTRIES LIST

Out[348]:

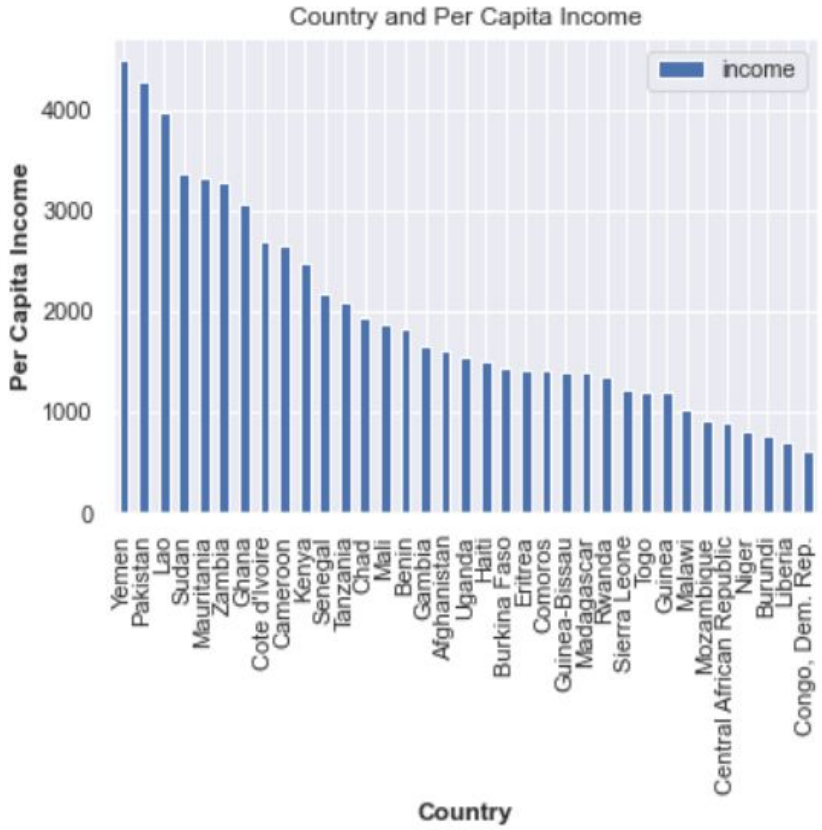
	country
0	Afghanistan
1	Benin
2	Burkina Faso
3	Burundi
4	Cameroon
5	Central African Republic
6	Chad
7	Comoros
8	Congo, Dem. Rep.
9	Cote d'Ivoire
10	Eritrea
11	Gambia
12	Ghana
13	Guinea
14	Guinea-Bissau
15	Haiti
16	Kenya
17	Lao
18	Liberia
19	Madagascar
20	Malawi
21	Mali
22	Mauritania
23	Mozambique
24	Niger
25	Pakistan
26	Rwanda
27	Senegal
28	Sierra Leone
29	Sudan
30	Tanzania
31	Togo
32	Uganda
33	Yemen
34	Zambia

Name: country, dtype: object

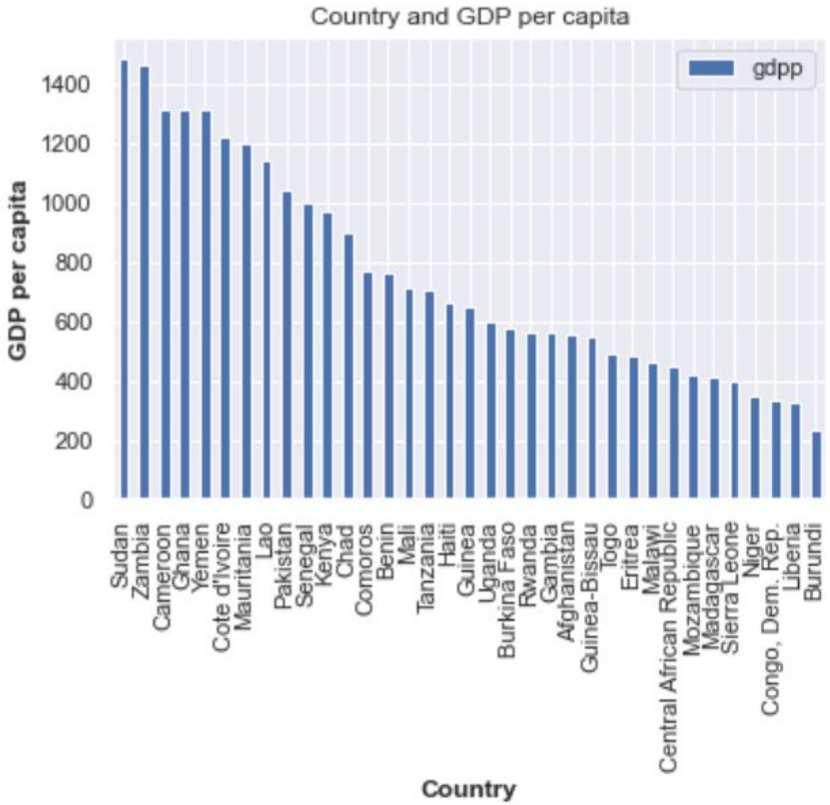
BarPlot for Child Mortality of countries which are in need of aid



BarPlot for Per Capita Income of countries which are in need of aid



BarPlot for GDP Per Capita of countries which are in need of aid



Final Observation

I found that the factor such as gdpp, Income and mortality rate play vital role in declaring any country as developed or underdeveloped. So above using two of the unsupervised clustering model i.e K-Means and Hierarchical clustering I got the names of countries that they have dire need of aid.

And so I plot the histogram according to GDP per capita , Income per capita and child mortality from the chosen country so that it can be selected according to need.

THANKS !